Knowledge Engineering and Data Science (KEDS)
Vol 7, No 2, December 2024, pp. 176–186

pISSN 2597-4602
eISSN 2597-4637

Optimal Strategy for Handling Unbalanced Medical Datasets: Performance Evaluation of K-NN Algorithm Using Sampling Techniques

Yulita Salim [a,1], Aulia Putri Utami [a,2*], Abdul Rachman Manga' [a,3], Huzain Azis [a,4], Fadhila Tangguh Admojo [b,5]

[a] *Universitas Muslim Indonesia*
Jl. Urip Sumoharjo No.km.5, Makassar 90231, Indonesia
[b] *University Kuala Lumpur Malaysia*
Jln Sultan Ismail, Bandar Wawasan, Kuala Lumpur 50250, Malaysia
[1] *yulita.salim@umi.ac.id;* [2] *auliaputriutami.iclabs@umi.ac.id*;* [3] *abdulrachman.manga@umi.ac.id;* [4] *huzain.azis@umi.ac.id;*
[5] *fadhila.tangguh@s.unikl.edu.my*
** corresponding author*

ARTICLE INFO

Article history:
Received 09 September 2024
Revised 12 November
Accepted 17 December 2024
Published online 24 December 2024

Keywords:
Oversampling
Undersampling
K-NN Algorithm
Medical image classification
Metrics evaluation

ABSTRACT

This study addresses the critical role of medical image classification in enhancing healthcare effectiveness and tackling the challenges of imbalanced medical datasets. It focuses on optimizing classification performance by integrating Canny edge detection for segmentation and Hu-moment feature extraction and applying oversampling and undersampling techniques. Five diverse medical datasets were utilized, covering Alzheimer's and Parkinson's diseases, COVID-19, brain tumours, and lung cancer. The K-Nearest Neighbors (K-NN) algorithm was implemented to enhance classification accuracy, aiming to develop a more robust framework for medical image analysis. The evaluation, conducted using cross-validation, demonstrated notable improvements in key metrics. Specifically, oversampling significantly enhanced lung cancer detection accuracy, while undersampling contributed to balanced performance gains in the COVID-19 class. Metrics, including accuracy, precision, recall, and F1-score, provided insights into the model's effectiveness. These findings highlight the positive impact of data balancing techniques on K-NN performance in imbalanced medical image classification. Continued research is essential to refine these techniques and improve medical diagnostics.

I. Introduction

In modern healthcare, medical image classification is vital in assisting clinicians with accurate and timely diagnoses [1]. The complexity and diversity of medical datasets, which include conditions such as Alzheimer's, Parkinson's Disease, COVID-19, brain tumours, and lung cancer, underscore the need for robust classification methods [2][3][4]. Despite significant advancements, a major challenge persists: addressing the issue of unbalanced datasets, which is critical for ensuring reliable and precise medical image classification [5].

Current literature indicates that imbalanced medical data can lead to biased model performance, often resulting in lower accuracy for underrepresented classes [6]. Addressing this gap requires a strategy that combines effective feature extraction methods with class imbalance handling techniques [7][8][9]. While techniques like Synthetic Sampling, Cost-Sensitive Learning, and Ensemble Methods have been explored, there remains a need for approaches that seamlessly integrate data balancing with feature extraction, particularly for multi-class medical image datasets [1][8].

This study addresses this gap by implementing a novel combination of Canny edge detection, Hu-moment-based feature extraction, and oversampling/undersampling techniques. Canny segmentation was selected for its precision in isolating key image structures. It is critical in medical contexts where feature clarity can impact classification outcomes [9][10][11][12]. Hu-moment features capture shape and texture, effectively representing medical images, while oversampling and undersampling address

https://doi.org/10.17977/um018v7i22024p176-186

data imbalance by ensuring a more equitable distribution across classes [13]. Together, these methods are expected to enhance model performance by addressing feature and data-related challenges in medical image classification.

The novelty of this approach lies in the integrated use of these techniques within a K-Nearest Neighbors (K-NN) framework. Unlike traditional methods, this approach leverages Canny segmentation and Hu-moment feature extraction to improve feature representation before applying class balancing. This integration will provide a more accurate and balanced classification framework for imbalanced medical image data. This research contributes a new perspective to developing medical image classification systems by optimising classification accuracy and reliability. It highlights the potential of combining these techniques for improved diagnostic tools.

II. Method

This research utilizes an experimental design to improve medical image classification by integrating Canny segmentation, Hu-moment-based feature extraction, and the K-Nearest Neighbors (K-NN) algorithm. This approach comprehensively explores the effects of each preprocessing and resampling step on model performance, as shown in Figure 1.

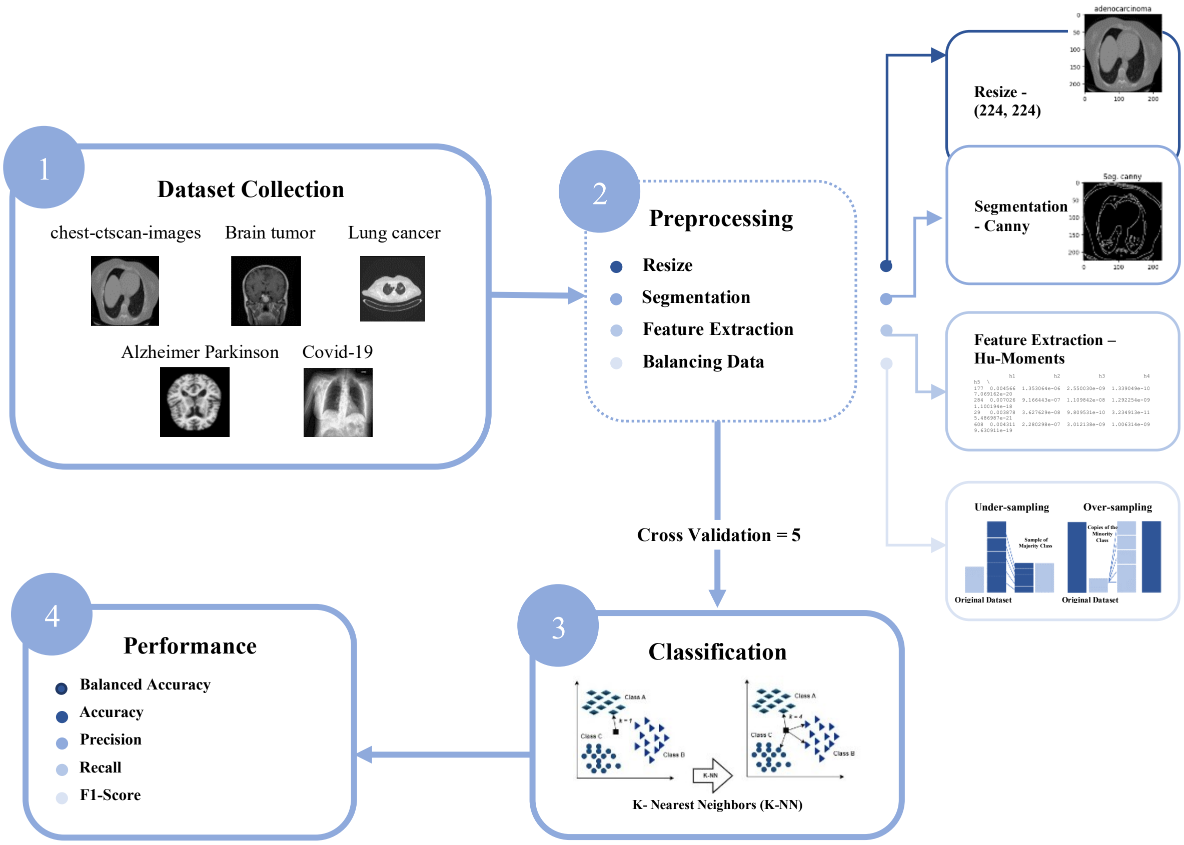


Fig. 1. Visualization of the research methodology flowchart

A. Collection of Medical Image Data

This research chose five distinct medical datasets to capture various diagnostic challenges. These datasets include images for Alzheimer's and Parkinson's diseases, COVID-19, brain tumours, chest CT scans, and lung cancer, all sourced from Kaggle. The selection of these datasets aims to create a comprehensive framework that reflects the diversity and complexity encountered in real-world medical applications. Each dataset includes important attributes, such as the number of cases, classes, and missing values, summarized in Table 1. This diversity enables a robust evaluation of the proposed classification approach across various medical conditions [6][14].

Table 1. Dataset information

Dataset	Number of Cases	Number of Attributes	Number: Name of Class	Number: Number in Each Class	Attribute Characteristics	Missing Values
chest-ctscan	613	7	4	195 115 148 155	Numeric	No
Brain Tumor Classification (MRI)	2870	7	4	826 822 395 827	Numeric	No
IQ- OTH/N CC - Lung Cancer	1097	7	3	120 561 416	Numeric	No
Alzheimer Parkinson Diseases	6477	7	3	2561 3010 906	Numeric	No
Covid- 19	251	7	3	111 70 70	Numeric	No

B. Data Preprocessing

Data analysis begins with adjusting the dataset size and applying Canny Segmentation on medical images to identify edges and distinctive features. These features are then extracted using the Humoment method. The next step was class balancing on each dataset due to an imbalance in the number of classes (Figure 1). Data balancing uses two approaches, namely under-sampling and over-sampling. The K-NN algorithm is used for classification, and this process is evaluated using a cross-validation method with K-fold 5 [14][15][16]. The final results were evaluated through performance measures such as accuracy, precision, recall, and f-measure.

The data preprocessing phase began by resizing all images to a standard dimension of 224x224 pixels. This step ensures uniformity across datasets, facilitating consistency in subsequent analysis and classification stages [17][18]. It also ensures that all images are uniform in size, providing benefits in subsequent stages of analysis or processing. The results of the data size adjustment process can be found in Figure 2.

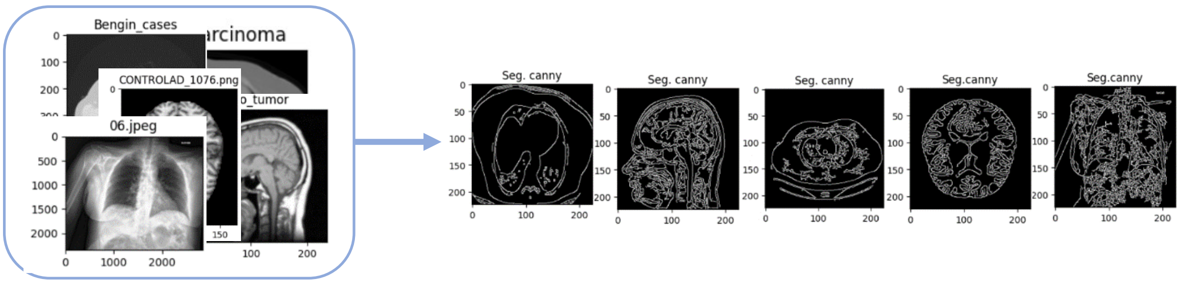


Fig. 2. Resizing dataset visualization

Next, Canny segmentation was applied to each image to detect edges, which is critical for highlighting structural features that help the classification model distinguish between classes. Canny segmentation calculates the gradient (G, G_x, G_y) to represent intensity changes, making edge features more prominent [19]. The main objective is to produce good edge detection while reducing the effects of noise, the equation as in (1).

$$G = \sqrt{G_x^2 + G_y^2} \tag{1}$$

G represents the image gradient at each pixel point, measuring how fast the image intensity changes around that point. , G_x Represents the gradient component in the horizontal direction $(x - axis)$, indicating how fast the intensity changes horizontally at that pixel point. G_y, on the other hand, is the gradient component in the vertical direction (y-axis), indicating how fast the intensity changes vertically at that pixel point. By combining G_x and G_y, we can understand the direction and degree of change in image intensity at each point. The result of the Canny segmentation process can be seen in Figure 3.

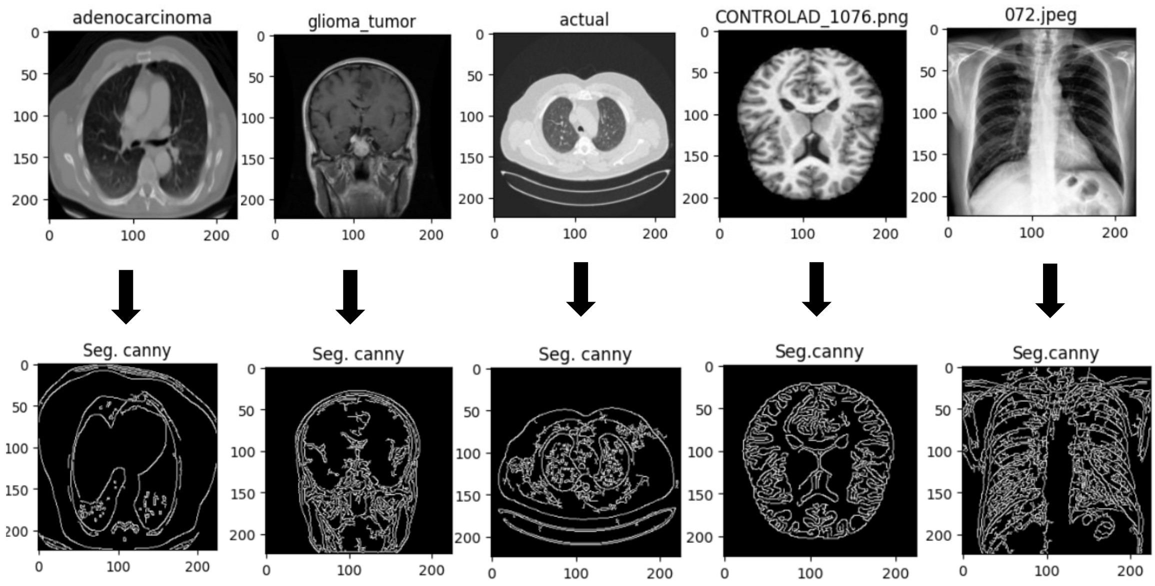


Fig. 3. Canny segmentation visualization

Following segmentation, Hu-moment-based feature extraction was employed to generate feature representations invariant to transformations in scale, rotation, and translation [19][20]. This technique calculates central and invariant moments (M_{ij}, h_{ij}), capturing unique patterns within each medical image. Figure 4 provides visualizations of the feature extraction results from one of the datasets. The moments are calculated by using the lift function of the image intensity distribution [21]. Some moments are mathematically transformed to produce invariant moments. The formula for Hu Moments can be seen as in (2) to (5).

$$h_{ij} = \frac{M_{ij}}{M_{00}^{(i+j)/2+1}} \tag{2}$$

Where h_{ij} is the i-th, jth invariant moment, M_{ij} is the i-th, jth central moment, and M_{00} is the zero-order central moment. The moments are calculated using the formulas as follows:

$$M_{pq} = \sum_x \sum_y x^p y^q I(x,y) \tag{3}$$

$$\bar{x} = \frac{M_{10}}{M_{00}}, \bar{y} = \frac{M_{01}}{M_{00}} \tag{4}$$

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q I(x,y) \tag{5}$$

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{(p+q)/2+1}} \tag{6}$$

With $I(x,y)$ is the image intensity at pixel point (x,y), $\bar{x}$ and $\bar{y}$ are the image centre of mass, μ_{pq} is the normalized central moment, and η_{pq} is the normalized invariant moment [22]. A scatter diagram and heat map showing the results of feature extraction using Hu-Moments on one of the datasets (Chest-ctscan) can be seen in Figure 4.

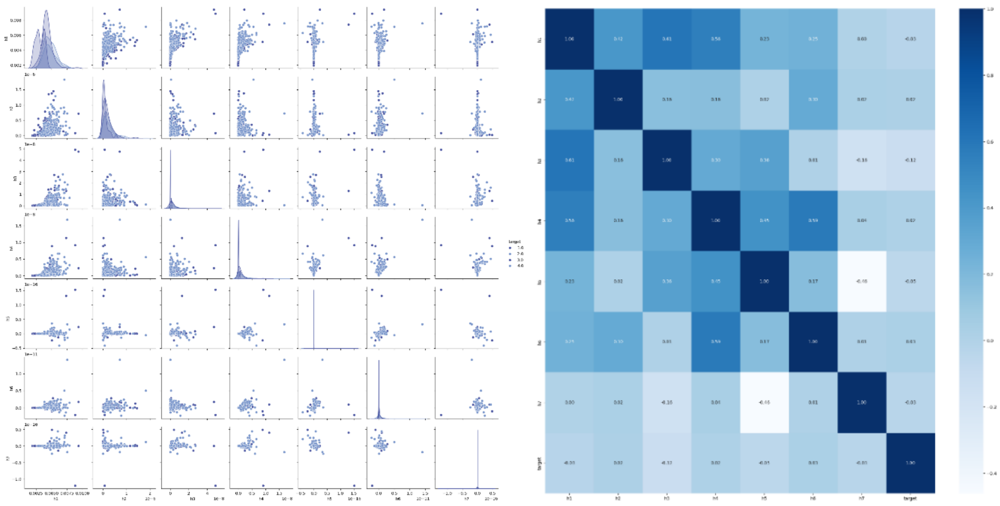
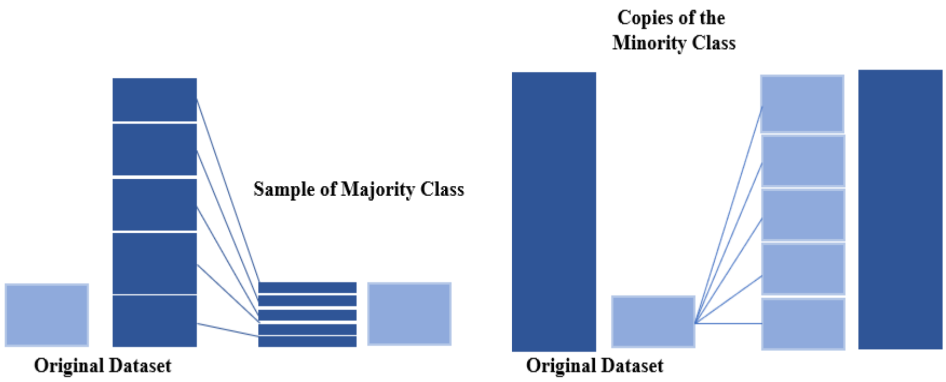


Fig. 4. Visualization of scatter plot and heatmap visualization for a chest-ctscan dataset using hu-moments feature extraction

Oversampling and undersampling techniques were applied to handle class imbalances present in these medical datasets. Oversampling enhances minority class representation by duplicating or synthesizing samples. At the same time, undersampling reduces the dominance of majority classes to promote balance. While oversampling benefits model sensitivity, undersampling helps prevent overfitting on majority classes. However, it may result in some data loss. Figure 5 illustrates the resampling process, and Table 2 details the balanced data distribution achieved for each dataset.



(a) Under-sampling (b) Over-sampling

Fig. 5. Data resampling visualization

Table 2. Data balancing

Datasets	Balancing data in class	
	Over-sampling	Under-sampling
Chest CT-Scan	195	115
	195	115
	195	115
	195	115
Brain Tumor classification (MRI)	827	395
	827	395
	827	395
	827	395
IQ- OTH/N CC -Lung Cancer	561	120
	561	120
	561	120
Alzheimer Parkinson Diseases	3010	906
	3010	906
	3010	906
Covid-19	111	70
	111	70
	111	70

In over-sampling, the sample size of the minority class is increased by adding copies of the existing sample or by creating a similar synthetic sample. The goal is to balance the majority and minority classes, hopefully reducing the risk of model bias towards the majority class. In contrast, under-sampling involves reducing the number of samples from the majority class, helping to address the class imbalance by reducing the dominance of the majority class and ensuring the model is more likely to learn patterns from the minority class.

C. Classification

Classification is a technique used to identify patterns or distinguishing features within a dataset, allowing the differentiation of each class [23]. This research uses classification to categorize each medical image into a specific diagnostic category by analyzing its features and patterns. The K-Nearest Neighbors (K-NN) algorithm was selected for classification, as it aligns well with the nature of medical datasets that include various conditions such as Alzheimer's and Parkinson's diseases, COVID-19, brain tumours, chest cancer, and lung cancer [24]. K-NN's simplicity and effectiveness in handling multi-class image classification make it a strong candidate for improving the accuracy and reliability of automated diagnosis systems [25].

The K-NN algorithm operates on the principle that objects with similar characteristics will likely belong to the same class. This approach is particularly relevant to medical imaging, where images with similar structural features often represent similar medical conditions. K-NN classifies each new image by measuring the distance between the image's feature vector and those of existing images in the dataset, determining the class based on the nearest neighbors [26][27]. The Euclidean distance formula, commonly used in K-NN, quantifies the similarity between feature vectors and is defined as in (7). By leveraging this distance-based classification, K-NN aids in accurately categorizing complex medical images, as illustrated in Figure 6.

$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{7}$$

Where $D(x, y)$ is the distance between two objects x and y, n is the number of features in each vector (n dimensions).

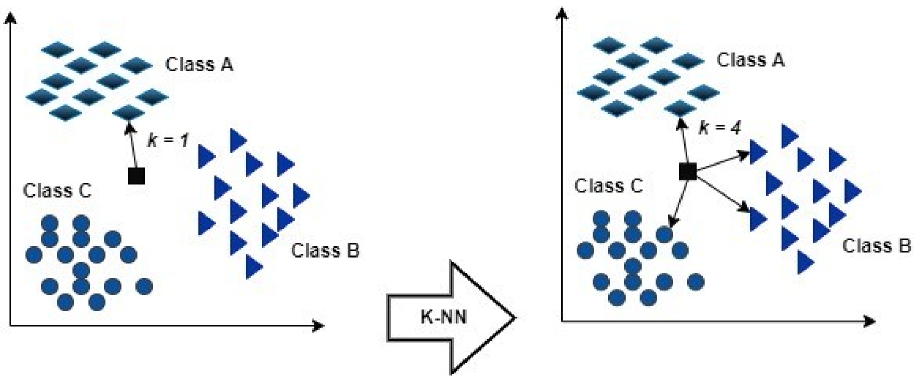


Fig. 6. K-NN algorithm

D. Evaluation Metrics

Evaluation metrics are measurement tools used to evaluate the performance of a model or system in handling classification tasks [28][29][30]. The classification model's performance was assessed using several evaluation metrics: balanced accuracy, precision, recall, F1-score, and specificity. Balanced accuracy combines true positive and true negative rates, providing a comprehensive measure that accounts for class imbalances. Precision evaluates the model's accuracy in identifying positive samples, while recall assesses its ability to capture all true positives. The F1 score balances precision and recall, particularly useful for datasets with uneven class distributions. Specificity measures the model's accuracy in identifying true negatives. These metrics collectively evaluate model performance to ensure a detailed analysis of reliability and diagnostic effectiveness across various medical conditions, with calculations as in (8) to (12).

$$Balanced\ Accuracy\ = \frac{1}{2}\frac{TP + TN}{(TP + FN)(TN + FP)} \quad (8)$$

$$Accuracy = \frac{TP + TN}{TP + TN\ FP + FN} \quad (9)$$

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

$$F1 - Score = 2\ \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \quad (12)$$

III. Result and Discussion

The results comprehensively evaluate the K-NN model's classification performance across diverse medical datasets, explicitly focusing on the impact of class imbalance handling techniques [31][32]. The evaluation framework considers three distinct scenarios: (1) the original, unbalanced dataset, (2) datasets subjected to oversampling to enhance minority class representation, and (3) datasets processed through undersampling to mitigate majority class dominance. The detailed performance metrics, including Balanced Accuracy, Accuracy, Precision, Recall, and F1-Score, are presented in Table 3 to Table 5, offering an in-depth comparative analysis.

Table 3 presents a detailed performance evaluation of the K-NN model when applied to unbalanced medical datasets, emphasizing key classification metrics, including Balanced Accuracy, Accuracy, Precision, Recall, and F1-Score. These metrics provide insight into the model's effectiveness in distinguishing various medical conditions despite class distribution disparities. Among the examined conditions, Lung Cancer and COVID-19 exhibited the highest balanced accuracy, reaching 0.59, suggesting a relatively stable classification performance. Conversely, Brain Tumor classification yielded the lowest balanced accuracy at 0.33, underscoring the inherent difficulties of imbalanced datasets, mainly when dealing with underrepresented classes. This discrepancy highlights the model's tendency to favour majority classes, leading to suboptimal performance for conditions with fewer training instances. The results underscore the critical need for effective class balancing techniques to mitigate the impact of data imbalance, ensuring that minority classes receive adequate representation in the learning process.

Table 3. Performance results of the K-NN classification algorithm on the original dataset

Datasets	Algorithm	Evaluation Matrics				
		Balanced Accuracy	Accuracy	Precision	Recall	F1-Score
Chest CT-Scan	K-NN	0.45	0.48	0.45	0.48	0.45
Brain Tumor		0.33	0.35	0.35	0.35	0.34
Lung Cancer		0.59	0.80	0.75	0.80	0.76
Alzheimer Parkinson		0.38	0.44	0.44	0.44	0.43
Covid-19		0.59	0.59	0.61	0.59	0.58

Table 4 presents the comprehensive performance evaluation after applying oversampling, revealing substantial improvements across all key classification metrics in multiple datasets. This enhancement is particularly pronounced in classifying medical conditions such as Chest CT-Scan abnormalities, Brain Tumors, Lung Cancer, Alzheimer-Parkinson diseases, and COVID-19 cases. Implementing oversampling effectively mitigated the challenges of class imbalance, leading to a marked increase in balanced accuracy. Notably, the Lung Cancer dataset exhibited the most significant improvement, with balanced accuracy surging from 0.59 in its unbalanced state to an impressive 0.80 after oversampling. This substantial gain highlights the pivotal role of oversampling in amplifying the model's capacity to detect minority class instances more effectively. The results underscore the efficacy of this approach in enhancing the K-Nearest Neighbors (K-NN) algorithm's discriminatory power, thereby improving overall classification robustness and ensuring more equitable recognition of underrepresented medical conditions. Such findings reaffirm the necessity of employing data-balancing techniques in medical image classification to bolster diagnostic reliability and optimize predictive performance.

Table 4. Performance results of K-NN classification algorithm on the dataset with over-sampling technique

Datasets	Algorithm	Evaluation Matrics				
		Balanced Accuracy	Accuracy	Precision	Recall	F1-Score
Chest CT-Scan	K-NN	0.54	0.54	0.55	0.54	0.53
Brain Tumor		0.40	0.39	0.41	0.39	0.38
Lung Cancer		0.80	0.80	0.82	0.80	0.79
Alzheimer Parkinson		0.59	0.59	0.60	0.59	0.58
Covid-19		0.61	0.62	0.63	0.62	0.61

Table 5 presents the detailed evaluation results following the implementation of the undersampling technique. While undersampling effectively balanced the class distributions within the datasets, it introduced inevitable trade-offs, particularly compared to the performance observed with oversampling. Notably, a decline in several key metrics was observed, especially within the Chest CT-Scan and Brain Tumor datasets, highlighting the impact of data reduction on classification performance. For instance, as depicted in Figure 7, the balanced accuracy for the Chest CT-Scan dataset decreased to 0.46, a notable drop from the 0.54 recorded in the oversampled scenario. This decline underscores the potential limitations of undersampling, mainly when dealing with datasets where information loss could adversely affect classification accuracy.

Table 5. Performance results of K-NN classification algorithm on the dataset with under-sampling technique

Datasets	Algorithm	Evaluation Matrics				
		Balanced Accuracy	Accuracy	Precision	Recall	F1-Score
Chest CT-Scan	K-NN	0.46	0.47	0.45	0.47	0.45
Brain Tumor		0.34	0.34	0.35	0.34	0.34
Lung Cancer		0.67	0.68	0.68	0.68	0.67
Alzheimer Parkinson		0.47	0.47	0.46	0.47	0.46
Covid-19		0.67	0.67	0.71	0.67	0.67

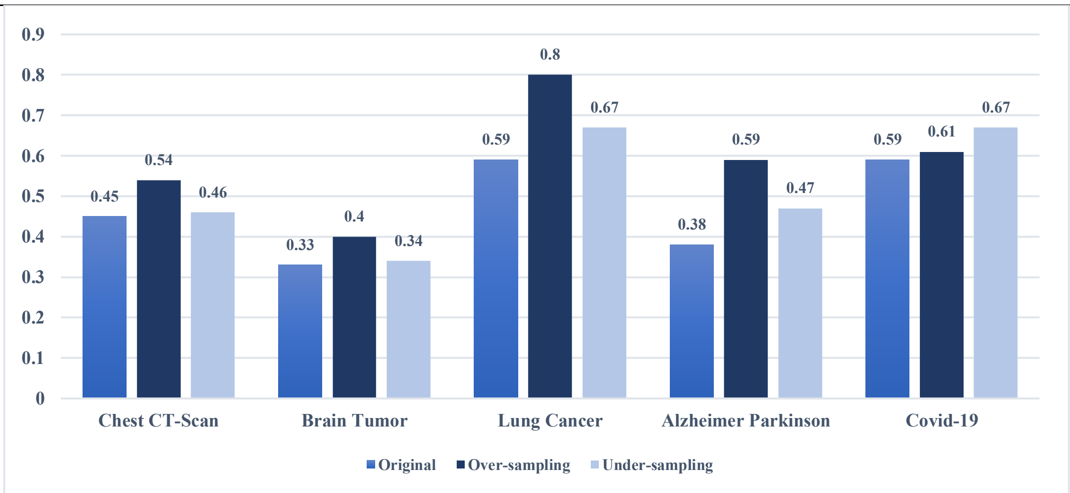


Fig. 7. Visualization of balanced accuracy performance

Conversely, the application of undersampling yielded a performance enhancement in specific datasets. Specifically, in the COVID-19 dataset, balanced accuracy improved, rising from 0.61 under oversampling to 0.67 with undersampling. This suggests that in cases where majority-class dominance leads to skewed decision boundaries, undersampling may enhance the model's ability to generalize across classes, ultimately improving classification reliability for minority-class instances. These findings emphasize the necessity of dataset-specific strategies when employing sampling techniques, as the effectiveness of undersampling is highly dependent on the nature and distribution of data. Therefore, while undersampling proves beneficial in mitigating class imbalance in some cases, its suitability must be carefully evaluated to ensure minimal loss of critical diagnostic information in medical image classification tasks.

The findings underscore the critical impact of class imbalance on the performance of the K-NN model across diverse medical datasets. Without balancing techniques, the model exhibited substantial variability in classification effectiveness, underscoring the necessity of addressing class disparities to ensure reliable predictions. Among the tested methods, oversampling generally led to superior performance across most datasets by amplifying minority class representation, thereby enhancing model sensitivity. In contrast, while undersampling effectively mitigated class imbalance by reducing the dominance of majority classes, it also introduced performance trade-offs, with specific datasets experiencing a decline in classification accuracy due to potential information loss. This was particularly evident in the Chest CT-Scan and Brain Tumor datasets, where undersampling resulted in lower balanced accuracy than the oversampled scenario.

Notably, the impact of these techniques was dataset-dependent, revealing that a one-size-fits-all approach to class balancing may not be optimal in medical image classification. The Lung Cancer dataset demonstrated the most significant improvement in balanced accuracy following oversampling, suggesting that generating synthetic samples can effectively improve model generalization for datasets with pronounced class imbalance. Conversely, the COVID-19 dataset exhibited better performance with undersampling, implying that reducing data redundancy may refine decision boundaries and improve classification robustness in cases where the majority class is disproportionately large. These findings highlight the necessity of tailoring resampling strategies to the unique characteristics of each medical dataset rather than relying on conventional balancing techniques in a generalized manner.

The implications of these results extend beyond technical performance, offering critical insights into the design of automated diagnostic tools. Ineffective class balancing can lead to biased predictions, potentially compromising clinical decision-making and patient outcomes. Therefore, selecting an appropriate data-balancing strategy is not merely an optimization task but a crucial component of ethical AI deployment in healthcare. Future research should explore advanced resampling techniques, such as hybrid approaches that combine oversampling and undersampling dynamically based on dataset properties. Additionally, refining hyperparameter configurations and integrating deep learning architectures—such as convolutional neural networks (CNNs) or attention-based mechanisms—could further enhance the adaptability and robustness of classification models. A more nuanced exploration of domain-specific balancing techniques, including cost-sensitive learning or synthetic augmentation strategies tailored to medical imaging, could significantly contribute to developing AI-driven diagnostic systems with improved clinical reliability and generalizability.

IV. Conclusion

This study analyzes how oversampling and undersampling affect the K-NN algorithm's performance in classifying imbalanced medical datasets. The findings reveal that oversampling generally improves balanced accuracy across most datasets, particularly in lung cancer detection, demonstrating its effectiveness in addressing class imbalance. At the same time, undersampling showed benefits for specific minority classes, notably enhancing balanced accuracy in COVID-19 detection. This research advances understanding in the field by offering insights into the nuanced effects of oversampling and undersampling on K-NN performance, contributing to the broader domain of automated medical diagnosis by guiding the selection of appropriate balancing techniques to improve classification accuracy for complex medical datasets. Recommendations for future research include exploring optimal parameter settings for these methods, considering alternative classification algorithms, and validating the model on more extensive and diverse datasets to strengthen automated diagnostic systems in real-world clinical contexts. This study thus lays a foundation for developing more robust, accurate classification models to support advancements in automated medical diagnosis.

Declarations

Author contribution

All authors contributed equally as the main contributor of this paper. All authors read and approved the final paper.

Conflict of interest

The authors declare no known conflict of financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We extend our deepest gratitude to the computer science faculty of Universitas Muslim Indonesia. Their guidance, expertise, and unwavering support have been key in realizing this research.

Additional information

Reprints and permission information are available at http://journal2.um.ac.id/index.php/keds.

Publisher's Note: Department of Electrical Engineering and Informatics - Universitas Negeri Malang remains neutral with regard to jurisdictional claims and institutional affiliations.

References

[1] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical image analysis using deep learning algorithms," Front. Public Heal., vol. 11, Nov. 2023.

[2] M. Khashei and N. Bakhtiarvand, "A novel discrete learning-based intelligent methodology for breast cancer classification purposes," Artif. Intell. Med., vol. 139, p. 102492, May 2023.

[3] B. M. de Andrade et al., "Grid Search Optimised Artificial Neural Network for Open Stope Stability Prediction," Chem. Rev., vol. 32, no. 2, pp. 600–617, 2020.

[4] T. Kumar, D. Kumar, and G. Singh, "Brain Tumour Classification Using Quantum Support Vector Machine Learning Algorithm," IETE J. Res., vol. 70, no. 5, pp. 4815–4828, May 2024.

[5] T. Huynh, A. Nibali, and Z. He, "Semi-supervised learning for medical image classification using imbalanced training data," Comput. Methods Programs Biomed., vol. 216, p. 106628, Apr. 2022.

[6] C. Tchito Tchapga et al., "Biomedical Image Classification in a Big Data Architecture Using Machine Learning Algorithms," J. Healthc. Eng., vol. 2021, pp. 1–11, May 2021.

[7] V. D. P. Jasti et al., "Computational Technique Based on Machine Learning and Image Processing for Medical Image Analysis of Breast Cancer Diagnosis," Secur. Commun. Networks, vol. 2022, pp. 1–7, Mar. 2022.

[8] S. P. Morozov et al., "MosMedData: data set of 1110 chest CT scans performed during the COVID-19 epidemic," Digit. Diagnostics, vol. 1, no. 1, pp. 49–59, Dec. 2020.

[9] N. Biswas, K. M. M. Uddin, S. T. Rikta, and S. K. Dey, "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," Healthc. Anal., vol. 2, no. October, p. 100116, 2022.

[10] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Deep Imbalanced Learning for Face Recognition and Attribute Prediction," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 11, pp. 2781–2794, Nov. 2020.

[11] H. Bakiler and S. Güney, "Estimation of Concentration Values of Different Gases Based on Long Short-Term Memory by Using Electronic Nose," Biomed. Signal Process. Control, vol. 69, p. 102908, Aug. 2021.

[12] G. Odongo, R. Musabe, and D. Hanyurwimfura, "A Multinomial DGA Classifier for Incipient Fault Detection in Oil-Impregnated Power Transformers," Algorithms, vol. 14, no. 4, p. 128, Apr. 2021.

[13] P. S. Singh, V. P. Singh, M. K. Pandey, and S. Karthikeyan, "Enhanced classification of hyperspectral images using improvised oversampling and undersampling techniques," Int. J. Inf. Technol., vol. 14, no. 1, pp. 389–396, 2022.

[14] N. Ahmed, A. Yigit, Z. Isik, and A. Alpkocak, "Identification of leukemia subtypes from microscopic images using convolutional neural network," Diagnostics, vol. 9, no. 3, 2019.

[15] A. T. Nagi, M. Javed Awan, R. Javed, and N. Ayesha, "A Comparison of Two-Stage Classifier Algorithm with Ensemble Techniques on Detection of Diabetic Retinopathy," 2021 1st Int. Conf. Artif. Intell. Data Anal. CAIDA 2021, no. April, pp. 212–215, 2021.

[16] K. Alpan, "Performance Evaluation of Classification Algorithms for Early Detection of Behavior Determinant Based Cervical Cancer," ISMSIT 2021 - 5th Int. Symp. Multidiscip. Stud. Innov. Technol. Proc., no. October 2021, pp. 706–710, 2021.

[17] M. Ozsagir, C. Erden, E. Bol, S. Sert, and A. Özocak, "Machine learning approaches for prediction of fine-grained soils liquefaction," Comput. Geotech., vol. 152, p. 105014, Dec. 2022.

[18] T. M. Fahrudin, P. A. Riyantoko, and K. M. Hindrayani, "Implementation of Big Data Analytics for Machine Learning Model Using Hadoop and Spark Environment on Resizing Iris Dataset," in 2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), 2022, pp. 429–434.

[19] E. A. Sekehravani, E. Babulak, and M. Masoodi, "Implementing canny edge detection algorithm for noisy image," Bull. Electr. Eng. Informatics, vol. 9, no. 4, pp. 1404–1410, 2020.

[20] H. Sharma and N. Kanwal, "Video interframe forgery detection: Classification, technique & new dataset," J. Comput. Secur., vol. 29, no. 5, pp. 531–550, Aug. 2021.

[21] H. Liu et al., "Lift-drag characteristics of S-shaped hydrofoil under different cloud cavitation conditions," Ocean Eng., vol. 278, p. 114374, Jun. 2023.

[22] L. Dai, G. Liu, L. Huang, G. Xiao, Z. Xu, and J. Ruan, "Feature transfer method for infrared and visible image fusion via fuzzy lifting scheme," Infrared Phys. Technol., vol. 114, p. 103621, May 2021.

[23] T. Whasphutthisit and W. Jitsakul, "Ensemble Learning Approach for Enhanced Road Deaths Prediction," in 2022 International Conference on Digital Government Technology and Innovation (DGTi-CON), Mar. 2022, pp. 15–19.

[24] M. Anila and G. Pradeepini, "Diagnosis of Parkinson's Disease Using Deep Neural Network Model," in 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Oct. 2021, pp. 1–7.

[25] S. Mishra et al., "Improving the Accuracy of Ensemble Machine Learning Classification Models Using a Novel Bit-Fusion Algorithm for Healthcare AI Systems," Front. Public Heal., vol. 10, May 2022.
[26] M. J. Mammoottil, L. J. Kulangara, A. S. Cherian, P. Mohandas, K. Hasikin, and M. Mahmud, "Detection of Breast Cancer from Five-View Thermal Images Using Convolutional Neural Networks," J. Healthc. Eng., vol. 2022, pp. 1–15, Feb. 2022.
[27] N. K. Kumar, G. S. Sindhu, D. K. Prashanthi, and A. S. Sulthana, "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers," 2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020, no. April, pp. 15–21, 2020.
[28] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," 2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020, no. April, pp. 243–248, 2020.
[29] P. Devan and N. Khare, "An efficient XGBoost–DNN-based classification model for network intrusion detection system," Neural Computing and Applications. Springer, 2020.
[30] H. Dickel, V. Podolskiy, and M. Gerndt, "Evaluation of Autoscaling Metrics for (stateful) IoT Gateways," in 2019 IEEE 12th Conference on Service-Oriented Computing and Applications (SOCA), Nov. 2019, pp. 17–24.
[31] A. N. Jaber, K. Moorthy, L. Machap, and S. Deris, "The importance of data classification using machine learning methods in microarray data," TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 19, no. 2, p. 491, Apr. 2021.
[32] M. S. B. A. Ghaffar et al., "Improving classification performance of four class FNIRS-BCI using Mel Frequency Cepstral Coefficients (MFCC)," Infrared Phys. Technol., vol. 112, p. 103589, Jan. 2021.